Supplementary Methods

Inference of the nucleotide substitution pattern

The coding sequences (CDSs) of SARS-CoV-2, RaTG13, and GX-Pangolin-CoV were downloaded from GenBank with the accession number NC_045512, MN996532, and MT040335, respectively. The two genomes of the GD-Pangolin-CoV coronavirus were downloaded from GISAID (EPI_ISL_410544) and Genome Warehouse (GWHABKW00000000) and merged to build the consensus sequence as previously described [1]. The CDS of GD-Pangolin-CoV consensus sequences were annotated with SARS-CoV-2 proteins using Exonerate v2.4 (--model protein2genome:bestfit --score 5 -g y) [2]. To avoid missing annotations in other coronaviruses, ORFs not included in GenBank were also annotated in this manner. To build a CDS alignment of SARS-CoV-2, RaTG13, GD-Pangolin-CoV, and GX-Pangolin-CoV, we first aligned the protein sequences of each gene using MUSCLE v3.8.31 [3] and then reverse translated them into codon alignments using RevTrans [4]. Subsequently, we concatenated all the aligned CDS sequences of each species. The sequence of the most recent common ancestor of SARS-CoV-2 and RaTG13 was inferred from the concatenated CDS sequences using CODEML in the PAML [5] package (**Fig. S1A**). The receptor binding domain (RBD) of S protein was masked due to a potential recent recombination [1]. The synonymous substitutions that lead to SARS-CoV-2 and RaTG13 were counted and summarized (**Fig. S1B**). Using CODEML in the PAML [5] package, we estimated that the SARS-CoV-2 CDS regions contained a total of 22,789 nonsynonymous and 6,443 synonymous sites, providing an approximate nonsynonymous/synonymous ratio of 3.5:1.

In silico simulations of molecular evolution with one outgroup

The simulation schemes are fully described in **Fig. 1A**. The simulations were performed in a Markov process in which mutation and selection occurred in a discrete time unit (a natural day). Previously, we observed a dN/dS ratio (ω, the ratio of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site) of ~0.05 between SARS-CoV-2 and RaTG3, suggesting that ~95% of the nonsynonymous mutations were removed by purifying selection during the evolution of these viruses [6]. Therefore, in the branch leading from N0 to N1 and N2, we assumed that both synonymous and nonsynonymous sites have the same mutation rate (u) and that synonymous mutations are neutral, but only 5% of the nonsynonymous mutations have a chance of preservation in each time unit during viral evolution. We set the initial genome (N0) to have the same codon composition as the concatenated CDS sequences of the SARS-CoV-2 reference sequence (NC_045512).

Each run of the simulation started from node N0, and for a day i ($i \geqslant 2$), new mutations randomly occurred in the sequence from the previous day (S_{i-1}) at a rate u, which had a Poisson distribution with a mean of 2.849×10^{-6} mutations/site/day. For a specific site at which that mutation occurred, the resulting nucleotide was generated using the mutational matrix shown in **Fig. S1**. Previously, we found that ~95% of the nonsynonymous mutations were removed by purifying selection during the evolution

of these viruses [6]. Therefore, in the branch leading from N0 to N1 and N2, we assumed that both synonymous and nonsynonymous sites have the same mutation rate (u) and that synonymous mutations are neutral, but only 5% of the nonsynonymous mutations have a chance of preservation in each time unit during viral evolution. Specifically, if a mutation was synonymous, it was preserved in the sequence S_i, and if a mutation was nonsynonymous, it was randomly preserved in the sequence S_i at a probability of 5%. The simulation was run independently in the branches that led to N1 and N2 on a daily basis. For the branch leading to N2, on day t_2, N3 and N4 evolved from N2 for t_3 days, except natural selection was not considered in the separation of N3 and N4. In the simulation, t_3 was fixed as 58 days for a θ of 0.1%, and 288 days for a θ of 0.5%. The MP method was used to infer the N2 sequences (N2') by comparing N3, N4, and N1; N2' was then compared with N2 (**Fig. 1B**). In each run of the simulation, the molecular evolution from N2 to N3 and N4 was repeated 1,000 times and the mean accuracy rate for ancestral state inference was calculated. In each run, both t_1 and t_2 started at 0 but stopped at a value ranging from 1,000 to 30,000 days (with an interval of 100 days). Each run of the simulation was repeated 200 times.

Simulations of molecular evolution with two outgroups

The simulation processes of molecular evolution with two outgroups (**Fig. S2**) were performed in a similar manner as described above. Specifically, we initiated the simulation from N5, which split into two branches, leading to N6 (resembling GD-Pangolin-CoV) and N0. After t_4 days, N0 further split into two branches, leading to N1 and N2, and after t_2 days, N2 further split into two branches, leading to N3 and N4. The N2 state was inferred based on the comparison of the nucleotides in N6, N1, N3, and N4 using the MP method, and the accuracy of ancestral inference was calculated by comparing the inferred N2' with the N2 state recorded in the simulations (**Fig. S2**). We set also the initial genome (N5) to have the same codon composition as the concatenated CDS sequences of the SARS-CoV-2 reference sequence (NC_045512). The other parameter settings for the simulations were the same as those described above.

The simulation was initiated at N5 on day 0 and was run independently in the branches that led to N0 and N6 on a daily basis. After t_4 days, N0 split into two branches, leading to N1 and N2, and after t_2 days, N2 split into two branches, leading to N4 and N5 for t_3 days. In the simulations, we fixed t_0 = 57,000 and t_4 = 27,000. The parameter settings of t_1, t_2, and t_3 were the same as those in the simulation using one outgroup.

Since the dS value between GD-Pangolin-CoV and SARS-CoV-2 is ~0.50 and that between RaTG13 and SARS-CoV-2 is ~0.17 [6], we estimated approximately t_0 = 110, 200, t_4 = 82,400 and $t_1 + t_2 + t_3$ = 27,800 days with a neutral substitution rate (u) of 2.849×10^{-6}/site/day (**Fig. S3**). Under such time settings in our simulations, the overall sequence similarity between N1 and N3 (or N4) was ~96%, which resembled the observed sequence similarity between RaTG13 and SARS-CoV-2. Moreover, the

genome similarity between N6 and N3 (or N4) was ~92%, which resembled the observed genome similarity between GD-Pangolin-CoV and SARS-CoV-2. Specifically, when using two outgroups, the uncertainty rate for ancestral inference at synonymous sites was 1.74% (95% CI, 1.71–1.76%) and that at nonsynonymous sites was 0.38% (95% CI, 0.37–0.39%).

Sequence divergence calculation

The sequence similarity was calculated between N1 and N3 and between N1 and N4, and the mean value was used to represent the similarity between the outgroup and the ingroup. The nucleotide substitution was not corrected in these processes.

Testing whether the performance of ancestral inference is affected by the strength of purifying selection among strains

Although purifying selection of nonsynonymous mutations was detected among SARS-CoV-2 strains [6-8], the strength may be weaker than that between SARS-CoV-2 and RaTG13. In the main TEXT, we assumed no purifying selection during the evolution of the branches leading from N2 to N3 or N4, in other words, we assumed no purifying selection between different strains of SARS-CoV-2. To test whether the performance of ancestral inference is affected by the strength of purifying selection among SARS-CoV-2 strains, we also conducted simulations with a dN/dS of 0.5 between two SARS-CoV-2 strains (*i.e.*, from N2 to N3 or N4) while keeping other parameters the same. **Fig. S4** shows that the accuracy of ancestral inference was hardly affected.

References

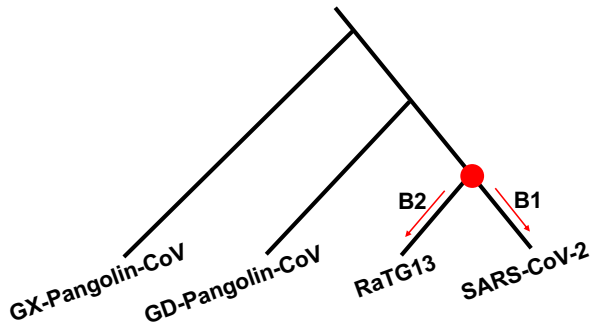
[1] Lam TT-Y, Shum MH-H, Zhu H-C*, et al.* Identifying sars-cov-2 related coronaviruses in malayan pangolins. Nature, 2020,
[2] Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics, 2005, 6: 31
[3] Edgar RC. Muscle: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, 2004, 32: 1792-1797
[4] Wernersson R, Pedersen AG. Revtrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res, 2003, 31: 3537-3539
[5] Yang Z. Paml 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol, 2007, 24: 1586-1591
[6] Tang X, Wu C, Li X*, et al.* On the origin and continuing evolution of sars-cov-2. National Science Review, 2020, 7: 1012-1023
[7] Shen Z, Xiao Y, Kang L*, et al.* Genomic diversity of sars-cov-2 in coronavirus disease 2019 patients. Clinical Infectious Diseases, 2020,
[8] Wang H, Pipes L, Nielsen R. Synonymous mutations and the molecular evolution of sars-cov-2 origins. bioRxiv, 2020, 2020.2004.2020.052019

Code availability

All simulations conducted under C++ scripts for the *in silico* evolutionary system are available at https://github.com/TaoLee0510/seq_sim

Supplementary Figures and legends

A



B

From	To	B1	B2	Average
A	C	0.075	0.066	0.0705
	G	0.729	0.693	0.711
	T	0.196	0.241	0.2185
C	A	0.082	0.056	0.069
	G	0.030	0.017	0.0235
	T	0.888	0.928	0.908
G	A	0.769	0.814	0.7915
	C	0.026	0.071	0.0485
	T	0.205	0.114	0.1595
T	A	0.128	0.094	0.111
	C	0.834	0.860	0.847
	G	0.037	0.045	0.041

Figure S1. The nucleotide substitution matrix in the synonymous sites.
(A) The phylogenetic tree of SARS-CoV-2, RaTG13, GD-Pangolin-CoV, and GX-Pangolin-CoV.
(B) The nucleotide substitution matrix in the branches from the most recent common ancestor of SARS-CoV-2 and RaTG13 (the red dot in Fig. S1A) to SARS-CoV-2 (B1 in Fig. S1A) and RaTG13 (B2 in Fig. S1A). The column "From" refers to the nucleotide in the synonymous sites of the most recent common ancestor. The column "To" refers to the nucleotide in the synonymous sites of SARS-CoV-2 or RaTG13. The occurrences in B1 and B2 were normalized, respectively, and the average values were used in the simulations of molecular evolution.

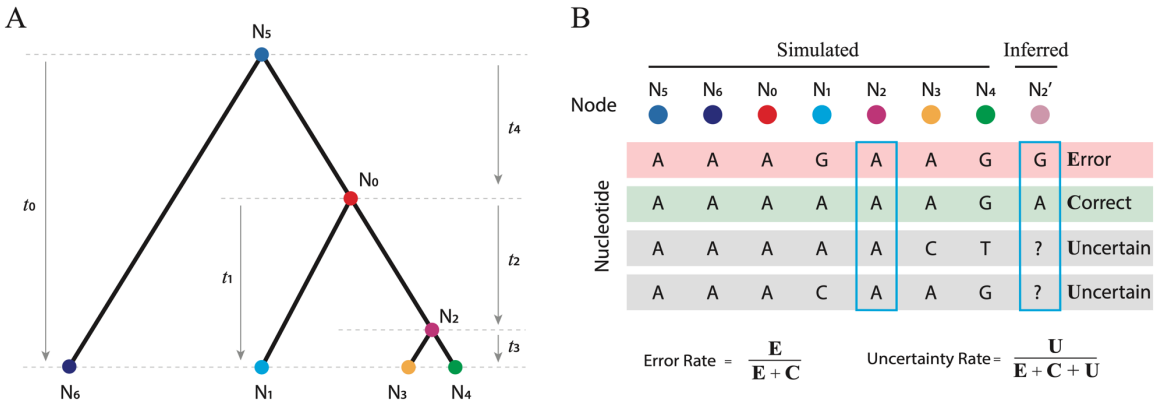


Figure S2. The scheme of molecular evolution simulation using two outgroups (N1 and N6). **(A)** The simulation was initiated from N5, which split into two branches, leading to N6 and N0. After t_4 days, N0 further split into two branches, leading to N1 and N2, and after t_2 days, N2 further split into two branches, leading to N3 and N4. **(B)** The N2 state was inferred based on the comparison of the nucleotides in N6, N1, N3, and N4 using the MP method, and the accuracy of ancestral inference was calculated by comparing the inferred N2' with the N2 state recorded in the simulations. See Fig. 1B for a more detailed description.

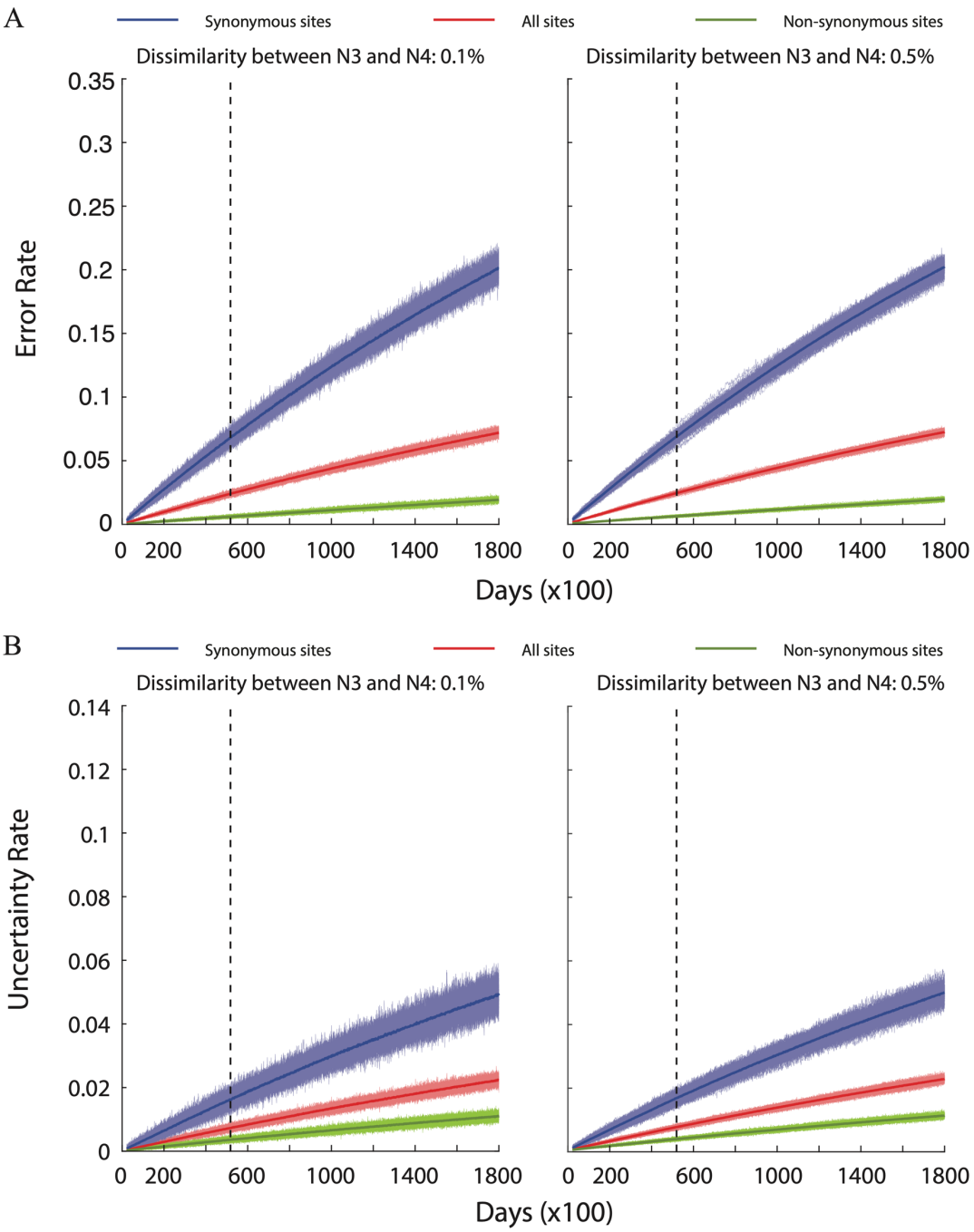


Figure S3. Results of molecular evolution simulation using two outgroups (N1 and N6). **(A)** The error rate for inferring the most recent common ancestor of N3 and N4 (*y axis*) increases as the divergence period (days) between N1 and N3/N4 increases (*x axis*). **(B)** The uncertainty rate for inferring the most recent common ancestor of N3 and N4 (*y axis*) increases as the divergence period (days) between N1 and N3/N4 increases (*x axis*). The left and right panels of (A) and (B) represent the results when the difference between N3 and N4 (θ) was 0.1% and 0.5%, respectively. The dashed lines represent the overall similarity equivalent to that between RaTG13 and SARS-CoV-2. The colors are the same as those in Figure 1.

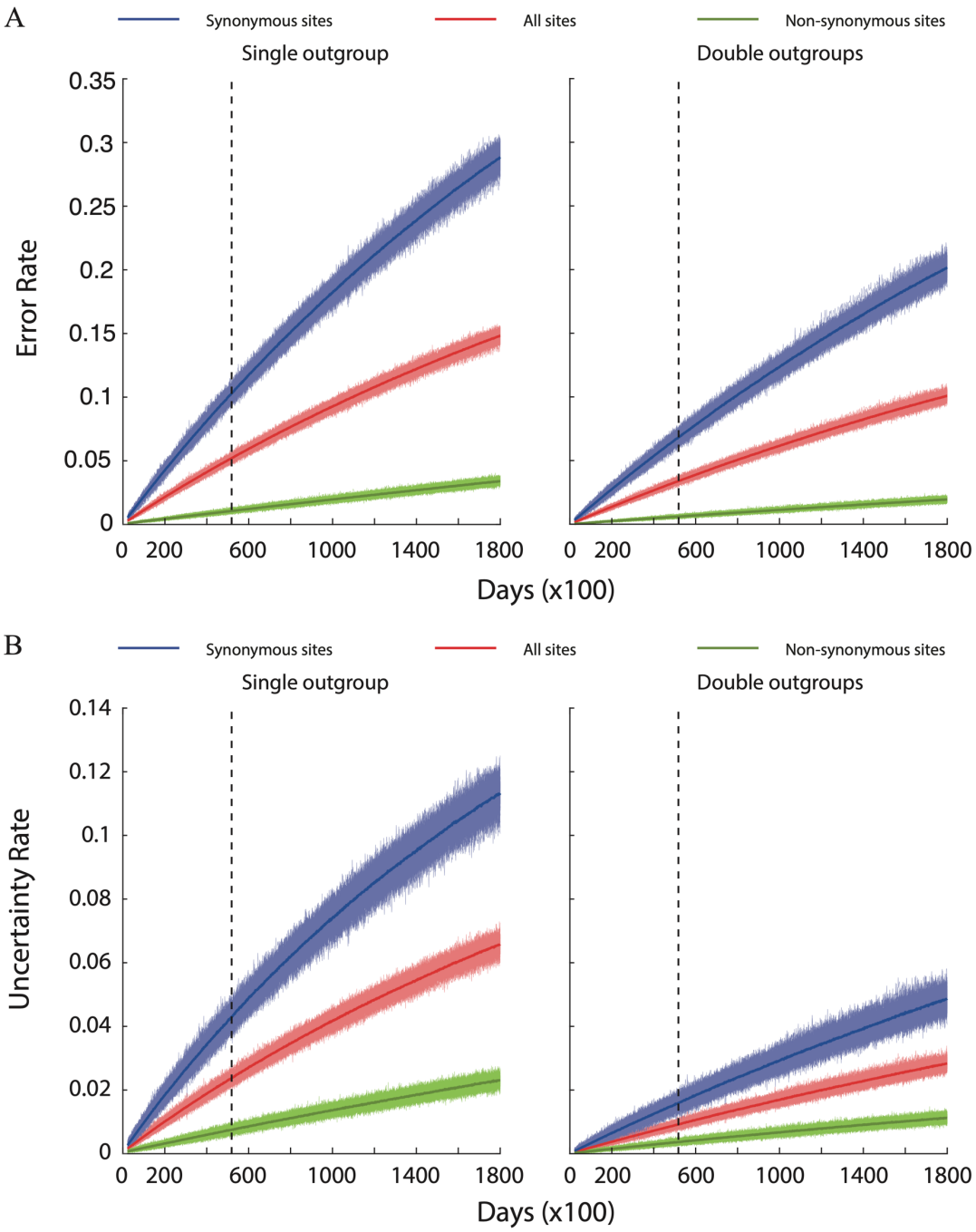


Figure S4. Results of molecular evolution simulation when dN/dS is 0.5 in the branches leading from N2 to N3 and N4.

(A) The error rate for inferring the most recent common ancestor of N3 and N4 (*y axis*) increases as the divergence period (days) between N1 and N3/N4 increases (*x axis*) using one (left) or two (right) outgroups. **(B)** The uncertainty rate for inferring the most recent common ancestor of N3 and N4 (*y axis*) increases as the divergence period (days) between N1 and N3/N4 increases (*x axis*) using one (left) or two (right) outgroups. The left and right panels of (A) and (B) represent the results when the difference between N3 and N4 (θ) was 0.1% and 0.5%, respectively. The dashed lines represent the overall similarity equivalent to that between RaTG13 and SARS-CoV-2. The colors are the same as those in Figure 1.